Optimal Shaping Codes for a TLC Flash Memory

Simeng Zheng, Andrew Tan, Carolina Fernández, Ismael González Valenzuela and Paul H. Siegel
Center for Memory and Recording Research, University of California, San Diego, La Jolla, CA 92093 U.S.A

***Abstract*—Shaping codes are distribution-matching codes that can be useful for coding over communication and storage channels with symbol costs and cost constraints. For example, the durability of a flash memory device can be quantified using wear costs associated with coded symbols, and in previous literature, shaping codes have been successfully applied to yield a notable lifetime gain for SLC (1 bit/cell) and MLC (2 bits/cell) flash devices. To account for the trending popularity of increased cell bit-density in flash, we formulate an optimal shaping scheme for TLC flash memories (3 bits/cell). The design procedure includes: wear cost estimation for 8 programmed levels, construction of an optimal shaping code using a concatenation of data compression and Varn coding to minimize the average wear cost with no overall expansion factor, and experimental performance evaluation on a 1x-nm TLC flash memory. Experimental results show a $10\times$ improvement in chip endurance.**

I. Introduction

In some applications, it is useful to model communications and storage channels as a costly channel, a variation of Shannon's discrete noiseless channel where output symbols and sequences of symbols are assigned a positive cost, where the meaning of cost depends on the application. For example, in DNA synthesis, the cost of a sequence of symbols is the number of synthesis cycles required to produce the sequence; in flash memory devices, the cost of a sequence of symbols is the amount of wear inflicted on the programmed cell. Since our goal is to improve the lifetime of TLC flash memory devices, we will view this problem as coding over a costly channel.

For a given finite alphabet $\mathcal{Y}$, any costly channel has a graph representation called a cost graph, which consists of a finite directed graph $G = (V, E)$, an edge labeling function $L : E \to \mathcal{Y}$, and a cost function $\tau : E \to \mathbf{R}^+$ such that there exists a path between any pair of states, the outgoing edges for a given state all have distinct labeling, and τ is non-negative and additive with respect to the edges. As a result, for a given initial state, any path $\gamma = e_1 e_2 ... e_k$ will correspond to the sequence $L(e_1)L(e_2)...L(e_k)$ and will have cost $\sum_{i=1}^{k} \tau(e_i)$.

Since the cost of a sequence directly measures its inflicted wear on a programmed cell of a flash device, the goal of a shaping code is to minimize the average cost per source bit for a given code expansion factor (i.e., inverse code rate). This corresponds to optimally shaping the probability distribution of sequences so that, roughly speaking, high-cost sequences have low probability and low-cost sequences have high probability.

The cost graph used in this study assumes symbol costs that are independent of context; that is, the cost is simply a function of the programmed cell level. Liu *et al.* [1] showed that, in this setting, for an i.i.d. source and a fixed expansion factor f, an optimal shaping scheme can be obtained by concatenating optimal data compression with a code from the compressed data that achieves the overall target expansion factor and minimizes the average cost per compressed bit. They also showed that a fixed-to-variable length Varn code [2] can be used in the second step to asymptotically achieve optimal shaping.

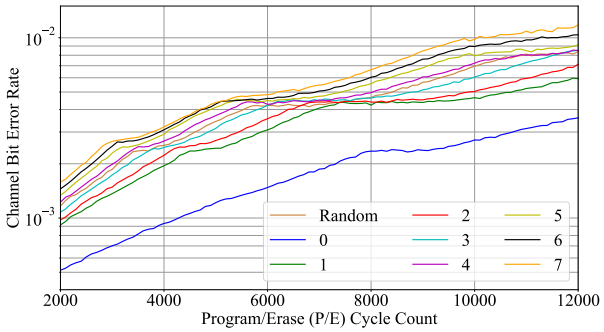

Fig. 1. Measured BER of pseudorandom data after inducing wear with data that is dominated by a single program level from P/E cycle 2000 to P/E cycle 12,000.

In [1], optimal shaping schemes with expansion factor $f = 1$ were implemented for a commercial 2y-nm MLC flash device Their performance was evaluated on an English language source (ASCII encoding) and a Chinese language source (UTF-16L encoding), and compared to scenarios with no coding, direct shaping coding [3], [4], and data compression alone. For the English language source, optimal shaping increased P/E cycling lifetime at a bit error rate (BER) of 1×10^{-3} by more than $2.6\times$ over no coding and more than $2.1\times$ over direct shaping. It also allowed the storage of more than $1.15\times$ the number of copies of the source than compression alone. For the Chinese language source, the corresponding gains were about $2.4\times$, $1.9\times$, and $1.25\times$.

In this project, we aim to design and implement optimal shaping codes for TLC flash memories. We give the construction and methodology in Section II and then present our experimental results in Section III. Extensions of the theoretical results in [1] to cost graphs where symbol costs are context-dependent are presented in [5].

II. Shaping Codes for TLC Flash

In triple level cell (TLC) NAND flash devices, three bits are stored in each cell to denote one of eight different program (i.e. voltage) levels. This means that the program level of a cell can be represented with an octal symbol, so we use $\mathcal{Y} = \{0, 1, 2, 3, 4, 5, 6, 7\}$ as the alphabet for our shaping code.

Based on [4], [6], for each $i \in \mathcal{Y}$ we assign symbol i a constant cost $c_i \geq 0$. Note that the corresponding cost graph will consist of a single state with eight edges, where each symbol and its associated cost is assigned to an edge.

The symbol costs c_i are estimated by performing repeated program and erase (P/E) operations on a TLC flash device and

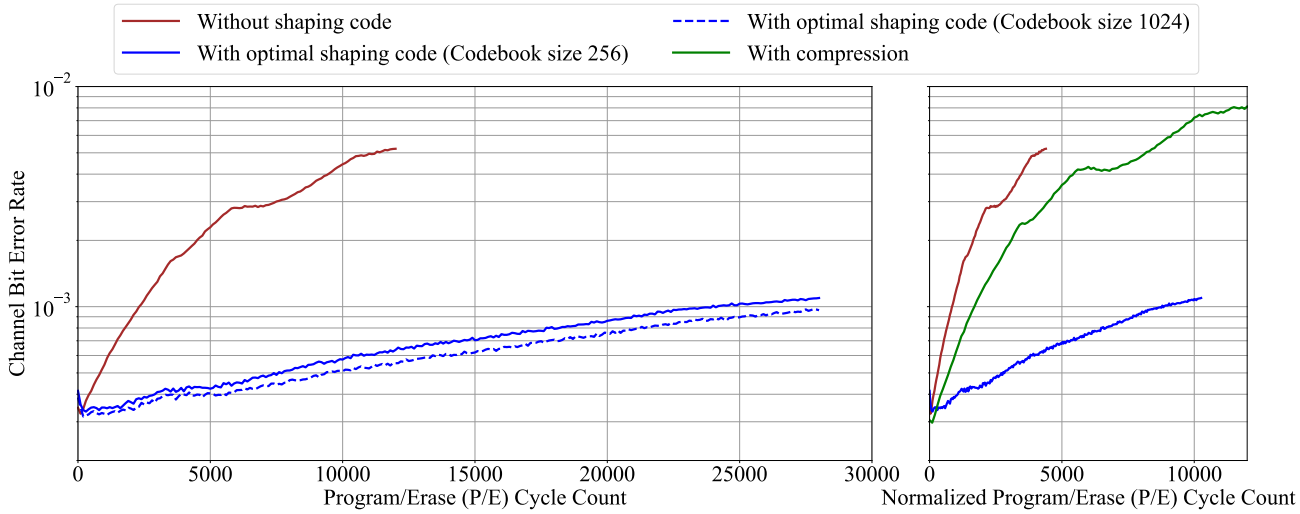

Fig. 2. (Left) Measured average channel BER comparison when all pages are programmed with Spanish novels data (brown solid), optimal shaping coded data with codebook size 256 (blue solid), optimal shaping coded data with codebook size 1024 (blue dashed) from P/E cycle 0 to P/E cycle 30,000. (Right) Measured average channel BER comparison in normalized P/E cycle count including compression data (green solid) from P/E cycle 0 to P/E cycle 12,000.

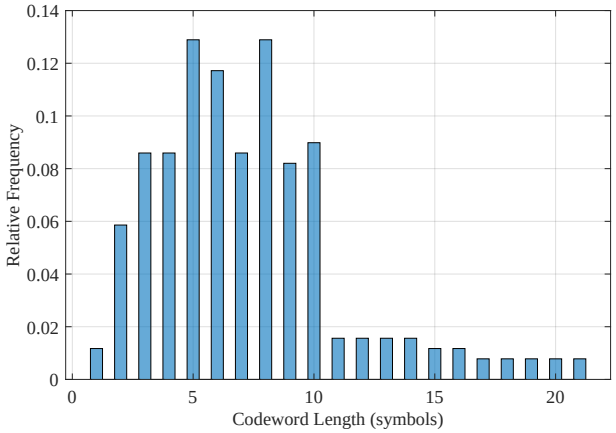

Fig. 3. Histogram of codeword length in optimal shaping codes with codebook size 256.

writing almost all cells with one certain program level. For every 100 P/E cycles, we estimate the induced wear of each program level by measuring the corresponding BER; the results are shown in Fig. 1, which also includes the BER of random data.

Let BER_{max} denote the maximum tolerable BER. Let T_R be the number of P/E cycles required to achieve BER_{max} when the chip is programmed with random data. For our experiment, we set $BER_{\text{max}} = 2 \times 10^{-3}$. For each $i \in \mathcal{Y}$, let T_i be the number of P/E cycles it takes to reach BER_{max} when inducing wear using only program level i. Then, the cost c_i associated with each level i is defined as $c_i = \frac{T_R}{T_i}$.

We utilize Theorem 2 in [1] to find the level probabilities p_i that minimize the average cost per stored level, $\sum_{i=0}^{7} p_i c_i$, for an overall expansion factor $f = 1$, where the entropy of the source, consisting of Spanish-language texts, is estimated from its compression factor (2.73) under LZ77 compression (gzip). We constructed two Varn codes of codebook sizes 256 and 1024. The symbol costs, c_i, for the 1x-nm TLC flash device, the optimal symbol probabilities, p_i, the relative contribution of each level to the total average cost, $p_i c_i$, and the empirical symbol probabilities achieved by the shaping codes, p_i^{256} and p_i^{1024}, are shown in Table I.

TABLE I
SYMBOL PROBABILITIES AND COSTS FOR OPTIMAL SHAPING CODES

Level	0	1	2	3	4	5	6	7
c_i	0.42	0.76	0.84	0.94	1.03	1.14	1.19	1.28
p_i	0.810	0.085	0.050	0.026	0.014	0.007	0.005	0.003
$p_i c_i$	0.340	0.065	0.042	0.024	0.014	0.008	0.006	0.004
p_i^{256}	0.777	0.110	0.065	0.025	0.097	0.054	0.049	0.032
p_i^{1024}	0.797	0.089	0.058	0.030	0.016	0.006	0.004	0.001

III. EXPERIMENTAL RESULTS

The left subfigure of Fig. 2 presents the BER performance of the original Spanish-language text and shaping-coded data. When channel BER is 1×10^{-3}, the shaping code with codebook size 256 (resp., 1024) achieves a lifetime gain of $10\times$ (resp., $11.7\times$) over uncoded data. In the right subfigure, we normalize the P/E cycle count by the compression ratio. We see that, compared to using compression alone, the shaping code with codebook size 256 increases the number of times the source text can be written before reaching a BER of 1×10^{-3} by a factor of about $5.2\times$. The histogram of codeword lengths in the Varn code of size 256 is shown in Fig. 3.

A notable observation is that the $10\times$ lifetime gain observed using a Varn code of size 256 for the TLC device is much larger than the $2.6\times$ gain achieved for the MLC device. This is a reflection of the level-dependent wear characteristics of the TLC device, as quantified in Fig. 1. It would be interesting to investigate whether an optimal shaping scheme could achieve even further lifetime gains as flash devices continue to scale up the bit density in each cell.

REFERENCES

[1] Y. Liu *et al.*, "Rate-constrained shaping codes for structured sources," *IEEE Trans. Inf. Theory*, vol. 66, no. 8, pp. 5261–5281, Aug. 2020.
[2] B. Varn, "Optimal variable length codes (arbitrary symbol cost and equal code word probability)," *Inf. Control*, vol. 19, no. 4, pp. 289–301, Nov. 1971.
[3] E. Sharon *et al.*, "Data shaping for improving endurance and reliability in sub-20 nm NAND," presented at the Flash Memory Summit, Santa Clara, CA, USA, Aug. 2014.
[4] Y. Liu *et al.*, "Shaping codes for structured data," in *Proc. IEEE Global Commun. Conf.*, Washington, DC, USA, Dec. 2016.
[5] Y. Liu *et al.*, "Rate-constrained shaping codes for finite-state channels with cost," in *Proc. IEEE Int. Symp. Inf. Theory*, Espoo, Finland, June 2022, pp. 1354–1359.
[6] Z. T. Blair, "Characterization of TLC flash memory," Master's Thesis, University of California San Diego, 2017.